DOCUMENT RESUME

ED 298 177 TM 012 339

AUTHOR McLarty, Joyce R.
TITLE Career Ladder Instrumentation: The Tennessee Experience.
PUB DATE Apr 87
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). Appendix A contains small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Career Ladders; Classroom Observation Techniques; Elementary Secondary Education; Interviews; Multiple Choice Tests; Professional Development; Questionnaires; *Teacher Evaluation; *Test Construction; Test Results; Test Use
IDENTIFIERS Teacher Testing; *Tennessee Career Ladder Program

ABSTRACT

The evaluation instruments used in the Tennessee Career Ladder program are described, and their evolution over the first three years of program implementation is traced. The instruments are designed to measure teacher performance and to determine outstanding teachers. They include: the Career Ladder Test of Professional Skills (a multiple-choice test), a classroom observation form, a candidate interview, a candidate portfolio and professional development and leadership summary, dialogues, principal interviews and questionnaires, student and peer questionnaires, and an evaluation consensus judgment. Effect sizes separating general education candidates who achieved Level I, II, and III status during 1985-86 for the 70 scores used in the Career Ladder evaluations are provided. In general, separation of the levels was excellent. Fifty of the scores (71%) separated Levels I and II by half a standard deviation or more, and 21 (30%) separated them by 0.8 standard deviations or more. Levels II and III were separated by at least half a standard deviation by 42 scores (60%) and by 0.8 standard deviation or more by 19 scores (27%). Appendix A contains data sources for the Career Ladder competencies and indicators. Appendix B presents analyses of scores, their differentiation among ladder levels, and the effect sizes of differentiation of levels. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298177

Career Ladder Instrumentation:

The Tennessee Experience

Symposium Presentation By

Dr. Joyce McLarty

Director
Test Development Department
ACT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joyce R. McLarty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

American Educational Research Association
Annual Meeting
April, 1987
Washington, D. C.

TM012339

BEST COPY AVAILABLE

Abstract

The evaluation instruments used in the Tennessee Career Ladder program are described and their evolution over the first three years of program implementation traced. Effect sizes separating general education candidates who achieved Level I, II, and III status during 1985-86 for the seventy scores used in the Career Ladder evaluations are provided.

In general, separation of the levels was excellent. Fifty of the scores (71%) separated Levels I and II by half a standard deviation or more, and 21 (30%) separated them by .8 standard deviations or more. Levels II and III were separated by at least half a standard deviation by 42 scores (60%) and by .8 standard deviation or more by 19 scores (27%).

Overview

In 1983, the Tennessee State Department of Education began a program to identify and reward Tennessee's best teachers. This Career Ladder program was designed primarily as a merit pay system. That is, teachers were to be rewarded for being outstanding teachers, not for performing extra duties. It was therefore necessary to define the characteristics of "outstanding teachers" and to develop ways of measuring them and basing decisions on the results of those measurements (setting standards). This paper provides an overview of the development of the definitions, measurements and standards, focusing on the various types of instruments used and evaluating their contributions to the measurement and decision processes.

Defining an Outstanding Teacher

The first step in developing the Career Ladder evaluation system was to define the objectives to be measured. These, together with the measurement methodology, would constitute the working definition of what Tennessee means when it identifies as teacher as meritorious. First a number of possible characteristics were selected from the research literature. These were reviewed, added to, and amended by Tennessee's teachers. Indeed, the first efforts resulted in hundreds of possible areas (objectives) for the evaluation of teachers.

Once this large set of objectives was assembled, it was reduced in three ways. First, those objectives which could not be reasonably measured in a statewide evaluation system were set aside for future consideration. Then objectives which were appropriate to only a few teachers or circumstances were eliminated. Finally, the remaining objectives were organized into a few, broad categories. The resulting domains of measurement were: Planning, Teaching Strategies, Evaluation, Classroom Management, and Leadership. These, and their subcategories (indicators of measurement), are provided in Appendix A.

Measuring the Objectives

The first step in measuring these objectives was to identify likely sources of information about the teacher's performance in each area. It was thought, for example, that the information regarding a teacher's ability to evaluate students might appropriately come from the teacher (via an interview or portfolio materials), the students (via interviews or questionnaires), the principal (via an interview or questionnaire), a written test, and possibly classroom observation. It was thought that peer teachers probably would not know very much about a teacher's

evaluation of students, and that the ability to identify this skill in the classroom observation would be somewhat limited.

For each area to be measured, as many possible sources of the required information as possible were identified. Existing instruments of all types (questionnaires, tests, observation schedules) were examined for their ability to provide information about the designated objectives. None was found to have an adequate level of content validity.

Accordingly, instruments were developed to measure the designated objectives by collecting data from each of the sources thought most likely to have the necessary information. The type of instrument used depended on the type of respondent (adults, students), the nature of the information required, and the cost/effectiveness of collecting the information in a reliable and objective way.
It is important to note that the emphasis was on collecting reliable, objective information about each teacher. This information would then be used to make judgments about the teacher's performance. Respondents were not asked to provide judgments directly.

For the first year of the program, seven instruments were developed. These and their current status are as follows:

Original instrument	Current status
1. Multiple-choice test	same
2. Classroom observation form	same
3. Structured candidate interview	Dialogue
4. Structured candidate portfolio	P. D. & L.
5. Structured principal interview	Prin. Questionnaire
6. Student questionnaire (primary, elementary and secondary forms)	same
7. Peer teacher questionnaire	(eliminated)

Appendix A indicates the objectives assessed with each instrument, originally and in the current configuration.

The instruments were pilot and field tested in spring of 1983 and all except the primary version of the student questionnaire were found acceptable. The primary student questionnaire, designed for students in Kindergarten through grade 2, was not sufficiently reliable and was later revised, retested, and finally accepted for inclusion in the evaluation system.

Setting Standards for the System

Once the instruments were developed and field tested, it was necessary to set standards for the evaluation system. Two

standards were needed, one for Level II (the lower of the two merit levels), and one for Level III (the higher level). It was required that the standards not be based on quotas (i.e. not be norm-referenced), and that they be determined and announced in advance of any actual evaluations.

Unfortunately, the complexity of the data made standard-setting in the ways most commonly used for multiple-choice tests impractical. Thousands of bits of information were available regarding each teacher's performance, many of them replicated (over occasions as with the observation, or over individuals as with the students). No clear criterion was available, either, although each of the three evaluators was asked to provide a direct judgment of each candidate and the average of these judgments was used as an index of the candidate's general level of performance for standard setting purposes.

After considerable discussion, it was decided by the Standard-Setting Panel to aggregate the information from each instrument over occasions, individuals, and (with some exceptions for the observation form) items addressing the same objective (domain and indicator). The scores so generated were placed on a stanine scale, weighted according to perceived importance and adequacy, and combined by domain. Distributions of scores for the field test candidates on the 80 scores originally created and on the five Domain scores were inspected and related to the average of the three evaluator judgments.

Standards were then selected for Level II and Level III for each of the five domains. Based on the Panel's recommendation, it was decided to require performance in each of the five domains at or exceeding the Level II cutoff score to qualify for Level II. For Level III, the candidate would be required to score at or above the Level III cutoff score in four of the five domains and at or above the Level II cutoff score in the remaining domain. This decision was based on the recognition that candidates in the field test had not performed consistently across the domains.

It was also decided that because none of the objective instruments could be expected to provide for all of the possible circumstances in which a teacher could demonstrate merit (or the lack thereof), the direct evaluator judgment would be included as an additional score. In order to provide as bias-free evaluation as possible, the three evaluators were to reach a consensus judgment about the candidate, and this judgment was to receive a weight in the decision in the same way as would any of the other scores. However, since the evaluators were providing judgments rather than information, the consensus judgments were made directly on the nine point decision scale rather than being transformed statistically to the stanine distribution.

Much of the discussion in the standard-setting meetings focused on the anticipated increase in scores on the actual

evaluation over the scores observed in the field test. Because of the extent to which score increases were expected to occur, and because the system was supposed to be "stringent", the standards selected were such that no field-test candidate would have qualified for Level III status. A more complete discussion of the original standard-setting effort can be found in McLarty (1986).

After the first year of implementation, scaling was changed from a stanine approach to a z-score one. Linearly transformed z-scores (constants were added for cosmetic purposes) were weighted and summed and the results expressed on a 200-800 scale. That reporting scale was the result of a common items equating process which fixed five scale points: the minimum (200), the qualifying score to be considered for Career Ladder II (450), the cutoff score for Level II (600), and cutoff score for Level III (700), and the maximum (800). All of the points were equated to the first year's equivalent scores in order to maintain standards despite some changes in items and instruments.

In addition to changing the scaling process and reporting scales, the second year of implementation the Career Ladder moved to the use of a single cut-off score, based on a weighted combination of the five domain scores, as the criterion. In order to compensate for the previous domain-by-domain qualification rules, minimum qualifying scores were set for each domain. It was necessary to achieve these scores in order to be considered for the higher levels of the Career Ladder.

In the third year of implementation, the same scoring process is being used as the second year except that no new equating is being done. Instead, no changes have been made in items or instruments, so the previous year's means, standard deviations, constants, and equating tables can be used.

The Instruments

Since the first year of implementation, 1983-84, a number of changes have been made in the Career Ladder instruments. These changes have been designed to improve the accuracy of the Career Ladder evaluation, reduce the burden on the candidates, and solve certain problems which occurred. Changes have also been made in the scoring of some instruments for similar reasons. Other changes have been considered but rejected or delayed for a variety of reasons. These are discussed here for each instrument individually.

A more complete discussion of the original instrumentation is available in McLarty, Furtwengler, and Malo (1985). This discussion provides a brief description of each instrument, the changes considered and implemented over the last three years, and data relating to its ability to separate general education

candidates who achieved Career Ladder Levels I, II, and III during 1985-86. The data cited here are presented in Appendix B.

Multiple Choice Test

The multiple choice test, titled the Career Ladder test of Professional Skills, was designed to assess the candidate's knowledge of the pedagogical aspects of planning, teaching strategies, evaluation, and classroom management. It has 100 items, 25 in each of the four domains measured. Each of these subtests is scored independently (no total score is used), scaled, weighted, and combined with the other scores used in the evaluation.

The Career Ladder Professional Skills test was difficult to construct because the state of the art in educational pedagogy was found to be lacking. It proved a problem to develop meaningful test items with a single correct response since no specific text or instructional program could reasonably be identified as the source of "right" answers. Because of the need to have legally defensible items under these circumstances, the test turned out to have a heavier emphasis on professional vocabulary than might have been wished. Items such as "If the teacher wished to employ negative reinforcement to correct this (described) problem, which of the following activities would be appropriate?" were seen by some candidates as trivial. However, it would have been extremely difficult to defend a single answer as the "best" one for the scenarios presented had the question been simply "What should the teacher do?".

Other problems with the test were the negative reactions of teachers to being required to take it at all, and the lack of a passing score. The latter led to a policy which allows teachers to retake the test whenever it is given and to use in evaluation only the highest score attained for each of its four subsections. While this policy is in line with encouraging teachers to learn the covered material and to perform as well as they possibly can, it causes two serious problems. First, teachers are under some pressure to retake the test until they achieve a perfect score on each subsection (since the higher the score, the more likely they are to qualify for the desired level). This leads to teachers studying for and retaking the test when they already have a score of, for example, 23 items correct of a possible 25. It is unlikely that the additional effort under these circumstances is warranted.

Second, this has resulted in the test becoming a less cost/effective way of assessing the teacher's knowledge base. The state is unable to afford the constant supply of new forms required by offering the test up to five times a year, and therefore the teachers are able to retake the same forms and increase their scores based on familiarity with the items as much as increased knowledge.

Despite the constant retakes, the test continues to differentiate the different levels of the Career Ladder. Based on

1985-86 data, the test differentiated Level I from Level II candidates on the order of .4 to .5 standard deviations, and Level II from Level III candidates on the order of .2 to .3 standard deviations. This should be interpreted in light of the weights of the individual subtest scores, in the final status decision, of 1.2% for Planning, 2.7% for Teaching Strategies, 1.2% for Evaluation, and 1.95% for Classroom Management.

Currently, the test is still required and utilized in its present format. There is some discussion of selecting a cutoff score to prevent the constant retakes which are undermining the validity of the scores and the cost/effectiveness of the program. Consideration has also been given to moving to a commercial test, such as the NTE. This would decrease program (but not candidate) costs at some sacrifice of content validity with respect to the objectives being measured. Inclusion of subject matter tests is also being discussed.

Classroom Observation Form

The classroom observation form is a three-part, relatively low-inference instrument, which utilizes multiple short "scans" in sections 1 and 2 to identify specific teacher and student behaviors. Section 3 provides for observation of those more static aspects of the classroom (for example, facilities and resources) which would not be expected to change during an hour-long observation. Evaluators are extensively trained to code the classroom events in strict accordance with established definitions, and are regularly checked on reliability and fidelity to the scoring protocol. Scoring is handled by computer programs; the evaluator does not provide judgment information on the observation form.

The observation form, because of its obvious objectivity, has been relatively popular with candidates. During the first year of its use, the evaluators appeared pleased with it as well. However after the first year, during which the observation form was considered a "secure" instrument and not made available to the candidates, the observation form was published in the Orientation Manual provided to all candidates. It is likely that providing the extensive detail inherent in the form has biased teacher activities in the classroom during the observation periods.

While the items initially chosen for inclusion were not a random sample of all possible teacher behaviors, it was assumed that they would generalize to a larger domain of behaviors of which these were an important but not sufficient component. Because the candidates now know exactly which behaviors to display this may no longer be the case.

Two other problems have been noted with the observation form. One is the difficulty of scoring it. Although several attempts have been made to generate complex, pattern-based

scores, these have not been successful. The form is currently scored primarily on a more-is-better basis, which may not be the most accurate approach for all situations. So far, however, this appears to be a technical problem rather than a defect in the instrument. Sufficient time to conduct the research necessary to identify and validate more sophisticated scoring approaches simply has not been available.

The second problem is the emphasis on quantity, as opposed to quality, inherent in the low inference approach. Although quality definitions can be included in the training of evaluators so that behaviors failing to meet the specified quality standard are not coded as having occurred, this is an awkward method of dealing with the situation.

Advantages of the low inference approach are the objectivity, the possibility of achieving high levels of reliability on the part of evaluators, and the level of detail available in the coded information for scoring and analysis.

The twenty-five scores generated from the observation form differed markedly in their ability to differentiate candidates who attained the different Career Ladder levels though most performed exceptionally well. The thirteen scores from section 1 separated Levels I and II all the way from not at all (the time a teacher required to get the class back on task after an interruption -- there were very few interruptions) to 1.2 standard deviations (the level of management of student behavior). The separation of Levels II and III was less variable, overall, ranging from about .1 standard deviations (the frequency with which the teacher was occupied with classroom procedures) to .8 standard deviations (the frequency of positive affect expressed by the teacher). Scores from section 1 of the observation were weighted from .18% (scores for use of classroom procedures and for teacher off task) to 3% (feedback provided to students) of the total evaluation score.

Only four scores are generated from section 2 of the observation. Two of the four separated Levels I and II well; percent of the class on-task (1.2 standard deviations), and the use of reasonable sanctions (.9 standard deviations). Both of these scores had weights of less than 1% in the decision. These scores separated Levels II and III by about .2 and .3 standard deviations, respectively. A third score, the degree to which the teacher gave correct information, correctly separated Levels II and III by a full standard deviation, and Levels I and II by about .1 standard deviation. This score exhibited strong ceiling effects, however. The fourth score, the degree to which the students expressed or exhibited lack of understanding separated Levels I and II by .1 and Levels II and III by about .3 standard deviations. These two scores were each weighted at .3% of the decision.

With the exception of the score used to detect potential race or sex bias (which was not expected to differentiate the

levels), the scores from section 3 of the observation form differentiated well between the Career Ladder levels. Differentiation between Levels I and II ranged from about .6 to 1.1 standard deviations, and between Levels II and III from about .5 to 1.0 standard deviations. Weights assigned to these scores ranged from .3% to 1.95%. Given the relatively small weights assigned to the individual observation-based scores, most appear to be contributing appropriately to the decision process.

An additional advantage of the observation instrument developed for use in Tennessee is that the teacher behavior portion is similar to a "zero-sum" approach. The coding is such that a teacher who spends excessive amounts of time on one activity may score high on that activity but will almost inevitably score lower on other activities for which no time remains. Spending all of one's time describing the content of the lesson (lecturing) will generally result in lower scores on monitoring student understanding and vice versa.

This appears to have resulted in teachers in all grade levels and subject areas being able to attain the higher levels of the Career Ladder despite different patterns of scores (more reviewing in the early grades, more description of content in the higher ones). When one scores lower in some areas, it is often because one is using the time in another area and as a result achieving higher scores in that one.

Structured Candidate Interview

The structured candidate interview has the dubious distinction of being retained for less time in the Career Ladder evaluation system then any other instrument. It didn't even make it through the first year.

The structured interview was conceptualized as being similar to those used in employment interviewing: specific questions, identical for all candidates, with keyed answers to be given credit (to maintain uniformity and objectivity). Some of the problems were obvious even before the field test: the first draft of the interview was estimated to take about eight hours! By removing large sections of questions, it was cut to an estimated one hour. However, in actual evaluations, two and three hours interviews were not uncommon, and longer ones occurred.

In addition, teachers did not like the structured nature of the interview. They wanted to be free to direct the discussion in directions of their own choosing. Many spent excessive periods of time preparing for the interviews (possibly to the detriment of their class preparations), and some experienced excessive anxiety about the interview situation.

Despite these difficulties, the interview process might well have survived had not the Tennessee Education Association

distributed the secure scoring criteria to teachers in order to "help them better prepare" for the experience. The scoring criteria were so much help, because of their highly objective nature, that interview results for candidates having access to the criteria were determined to be invalid. Since the individual candidates did not know that they were not supposed to have the criteria, and since it was not possible to accurately determine which candidates did and did not have the criteria, the structured interview was eliminated from the evaluation during the first year and the weight assigned to evaluator consensus judgment was increased from 20% to 25%. The results from those interviews which were conducted were never used.

For the second year of implementation, the structured interview and the portfolio were recombined into a new instrument called a Dialogue, which will be described below.

Structured Candidate Portfolio and
Professional Development and Leadership Summary

In many respects, the portfolio was similar to the interview. It, too, had problems with length and with the security of the scoring keys. Teachers spent incredible hours preparing portfolio materials in ways which were neither necessary nor desirable. While the original idea was simply to have the teacher present unit and lesson plans which the teacher had previously designed and found effective, many teachers created new materials and worked long and hard to present them in the most impressive formats possible.

The portfolio scoring keys were also released by the Tennessee Education Association, to which the Department of Education responded by reducing weight assigned to those scores resulting from the portfolio items identified as most likely to have been invalidated by knowledge of the scoring key. During the second and succeeding years of the Career Ladder, the portfolio was reduced to little more than a vita, in which the teacher specifies and describes professional growth activities relevant to the Leadership domain such as taking courses, conducting workshops, and assisting other teachers. The remaining aspects of the portfolio (unit and lesson plans, for example) have been combined with the Interview in the Dialogues.

The remaining aspects of the portfolio, called the Professional Development and Leadership summary (PDL for short) generated two scores in the Leadership domain, for professional development and for leadership. These are weighted at 2.4% and 2.6% of the decision, respectively, and each separated Level I from Level II and Level II from Level III by about half of a standard deviation.

Dialogues

The Dialogues are a restructuring of the Interview and portions of the Portfolio. In each of the three Dialogues, an evaluator meets with a candidate for one hour. The first half of the hour is structured, with a specific focus designated (the first Dialogue covers planning; the second, teaching strategies; and the third, evaluation). Teachers are provided with information on what will be covered, and are asked to bring specific materials (unit and lesson plans, for example) to discuss with the evaluator. Scoring is based on a qualitative key, with the essential considerations known to the candidate.

The ten scores generated from Dialogue information were quite consistent in their separation of Career Ladder candidates who attained the different levels. Separation of Levels I and II ranged from about .5 to .9 standard deviations, (four scores showed a separation of about .7 and three more about .8). Seven of the scores separated Levels II and III by about .8 standard deviations, and the remaining two by about .9. Weights assigned to the scores ranged from .7% to 2.4%.

The second half of the Dialogue is "candidate's choice", and the teacher may present any information or topic desired. This aspect is included by the evaluator in the consensus judgment. It is possible that the consistently strong separation of the Career Ladder levels observed in the Dialogue scores results in part from the influence of the dialogue experience on both the Dialogue score and the evaluator consensus score. Thus, the effective weight of the dialogue in the decision process is somewhat higher than the explicit weights would indicate.

Both evaluators and candidates have been positive about the contribution of the Dialogues to the Career Ladder evaluations.

Structured Principal Interview and Principal Questionnaire

The principal interview was designed as a structured interview in which the evaluator would ask set questions of the candidate's principal and record the responses. As with the other instruments, the principal interview was designed to collect information about the candidate's behavior rather than the principal's judgment regarding the candidate's appropriate Career Ladder status. It was felt that the principal would be more likely to provide correct and complete information to the evaluator directly, and might prefer to provide the information verbally.

A number of problems have occurred with the collection of information from principals. The first was that principals resented the structured questions asked of them by evaluators.

They indicated that it would be better to simply provide them with a questionnaire format, thus saving them and the evaluators time. This was done before the first implementation of the process.

The conversion to a stanine scale in scoring the principal questionnaire data resulted in some scores which were unanticipated by the principals and repugnant to them. Because of the wording of the items and response scale, and because of the extremely negatively skewed distributions encountered on the field test data for the principal questionnaires, principals who marked items on the scale assuming a particular outcome for the teacher often encountered a less positive outcome and an angry teacher.

For example, on an item asking the principal to indicate "How much of the time does this teacher keep the students on task?", a principal might have marked "3" on the 5 point scale to indicate average performance on this indicator. However, "3" was defined as "about half of the time" in the response key provided, and this, together with the tendency of many principals to ascribe near perfect performance to their teachers, resulted in what appeared to the principal to indicate average performance (3 on a 5 point scale) being scaled to indicate very poor performance (for example, a "3" on a 9 point stanine scale).

This combination of misunderstandings was, perhaps, the most serious problem after the first year of the Career Ladder program. (The most serious problem during the year having been the release of the secure rating scales for the interview and portfolio.) For the second year, the principal questionnaire was changed to permit principals to provide a direct judgment (much as in the evaluator consensus) on the final decision scale. This allows the principal, in effect, to vote on the final status of the teacher. While this change has not eliminated the extremely high scores assigned to most teachers by their principals, it has prevented the inadvertent "punishment" of those principals who provide less inflated ratings which occurred because their scores were scaled in with all of the other principal ratings.

Despite the severe negative skews, the five scores generated from the principal questionnaire separated teachers assigned to Level I and II by .5 to nearly .9 standard deviations. Separation of Level II and Level III was not as good, but the scores still separated those levels by about .3. Weights assigned to the principal's scores ranged from 1.5% to 3.0%.

It appears that principal input to the evaluation process is likely to continue and, perhaps, increase. There has been considerable discussion about increasing the emphasis on the appropriate evaluation of teachers in the administrators' Career Ladder system, and of having selected principals serve as Career Ladder evaluators for their teachers (currently, principals do not serve as evaluators in the teacher Career Ladder system). The

primary problems with this approach appear to be the difficulty (and lack of cost effectiveness) of having principals undergo the three to four weeks of extensive training (and periodic retraining) required to become reliable in the use of the current evaluation system, and the perceived importance of having evaluators be peer teachers from outside of the school system in which they are evaluating.

The possible advantages of additional principal involvement would be the long-term effectiveness of the investment in the training (principals would continue to evaluate teachers: peer teachers return to the classroom after one or two years as evaluators), the integration of Career Ladder evaluations with the regular evaluations of teachers, and a desire not to undermine the authority of the school principal (as the line supervisor for the teacher). It is likely that several approaches to increasing principal involvement in teacher evaluation will be pilot tested in the next year or two.

Student Questionnaires

The student questionnaires were designed to collect information on the teacher's teaching strategies, classroom management, and evaluation practices. As with the other instruments, they were intended to yield scorable information rather than judgments. For example, the student was asked to indicate how frequently "This teacher notices when I am not paying attention."

The elementary and secondary forms of the student questionnaires differ primarily in the vocabulary and syntax levels used for the items. Each generates eight scores. These scores vary in weight from .7% to 1.6%. Despite a significant negative skew, the questionnaire data differentiated the Career Ladder levels consistently. Levels I and II were separated by .4 to .8 standard deviations, and Levels II and III were separated by .4 to .7 standard deviations. The elementary form/students tended to separate the levels slightly better than the secondary ones.

The primary student questionnaire generates three scores; one each for teaching strategies, evaluation, and classroom management. These are weighted at 3.3%, 1.5% and 4.05% of the decision, respectively. The teaching strategies score reversed Levels I and II by -.2 standard deviations, and separated Levels II and III in the correct direction by .7 standard deviation. The evaluation score separated Level I and II by .8 and Level II and III by .3, while the classroom management score separated Levels I and II by a full standard deviation, but did not separate Levels II and III at all. Based on 193 primary teachers evaluated in 1985-86, only the separation of Levels II and III for the teaching strategies score and the separation of Levels I and II for the other two scores were statistically significant.

The student questionnaires tend to have high levels of internal consistency and fairly substantial negative skews. Because the primary version has only nine items, it has a somewhat lower level of reliability than the other instruments, and has been the topic of a continuous discussion throughout the implementation of the Career Ladder evaluation system. Many evaluators, candidates, and staff have recommended not using questionnaires with students this young, while others believe that these students are an important source of information about the teacher. Other methods of collecting this information from younger children have been discussed, as has the possibility of using parent information to replace this information source. In the meantime, the primary questionnaire continues in use.

Peer Questionnaire

The peer questionnaire was one of the most popular instruments, primarily because each teacher candidate was permitted to select six peers from whom three were randomly selected to respond to the questionnaire. The instrument covered only the Leadership domain. Like the other questionnaires, the peer questionnaire was designed to collect information rather than judgments. Peers might be asked, for example, how frequently this teacher shared materials and professional information with other teachers. The single score generated from this instrument was extremely negatively skewed, but nevertheless separated Level I and Level II candidates by .7 standard deviations and Level II and III candidates by .4 standard deviations.

Because of the consistently high scores and the difficulty, disruption and cost of locating the appropriate peers and getting them to complete the instrument, the peer questionnaire was eliminated after 1985-86.

Evaluator Consensus Judgment

After each of the three evaluators has spent a day with the candidate, the evaluators are brought together to discuss the candidate and to assign a consensus judgment. The consensus discussion focuses on the notes that the evaluators have taken during their visits. The other instruments and the scores based on them are not made available to the evaluators. This is because those scores are already included in the evaluation with specified weights. To include them a second time would result in overweighting their contributions.

Consensus decisions are made directly on the final decision scale. Four scores are generated, one for each domain except Leadership. Weights assigned relative to the status decision are 3.75% in Planning, 8.75% in Teaching Strategies, 3.75% in Evaluation, and 6.25% in Classroom Management. These

weights reflect both the importance of the domain in the total score, which varies, and the 25% contribution assigned to the evaluator within each domain (for example, the evaluator contributes 25% to the Planning domain, which is itself 15% of the total, resulting in an over-all contribution of 3.75%). The evaluator consensus scores separated Levels I and II by 1.3 to 1.5 standard deviations, and Levels II and III by .9 standard deviations (the same for each of the four scores).

In general, the evaluators believe that they can reliably distinguish different levels of teacher performance, and can detect a "canned" performance (one which is designed to yield a high score but is not typical of the teacher's normal classroom behavior). In general, the teachers like their evaluators personally, but remain distrustful of the consensus judgment as somehow less objective (or, possibly, more biased) than the other instruments.

Summary

In general, the instruments and scores used in the Tennessee Career Ladder evaluation system worked remarkably well for this group of general education candidates. Most of the seventy scores used in the evaluation were given weights of less than 2% in the total decision, and most had severely skewed distributions (questionnaires, negative; observations, positive). Despite these factors, the different objectives measured, and the different respondents completing the forms, the scores consistently separated the outcome candidate groups. The numbers of scores separating the groups, by the size of the separation in standard deviation units, were as follows:

Separation in standard.. deviations..	.Numbers of scores separating .Levels I & II	evels II & III
less than 0.0	2	0
0.00 - .09	2	1
.10 - .19	2	4
.20 - .29	3	6
.30 - .39	4	11
.40 - .49	7	6
.50 - .59	11	10
.60 - .69	13	8
.70 - .79	5	5
.80 - .89	11	13
.90 - .99	1	3
1.00 - 1.09	2	3
1.10 - 1.19	2	0
1.20 or greater	5	0

Separations this great would argue for substantial agreement among the data sources and across the instruments and respondents as to the "correct" categorization of the candidates. This, in turn, lends substantial credibility to the Career Ladder evaluation system.

References

McLarty, J. R. (April, 1986). Tennessee's Career Ladder: How far to space the rungs. Paper presented at the American Educational Research Association and National Council on Measurement in Education annual meeting, San Francisco, CA.

McLarty, J. R., Furtwengler, C., & Malo, G. (April, 1985). Using multiple data sources in teacher evaluation. Paper presented at the National Council on Measurement in Education annual meeting, Chicago, IL.

Appendix A

Data Sources for the Tennessee Career Ladder Competencies and Indicators for 1984-85 & 1985-86

1984-S

Data Sources for

COMPETENCIES AND INDICATORS

	Classroom Observation	Portfolio	Applicant Interview	Principal Interview	Peer Questionnaire	Student Questionnaire	Written Test
I. Prepares for instruction effectively.							X
A. Establishes appropriate instructional goals and related objectives consistent with the curriculum.		X	X	X			
B. Prepares instructional plans and materials incorporating principles of effective instruction.	X	X	X	X			
C. Creates, selects or modifies instructional plans and materials to accommodate learner instructional levels.	X	X	X	X			
II. Uses teaching strategies and procedures appropriate to the content, objectives, and learners.							X
A. Provides a clear description of the learning task and its content.	X	X				X	
B. Monitors learner understanding and reteaches as necessary.	X	X	X			X	
C. Provides learners appropriate practice and review.	X	X	X			X	
D. Establishes and maintains learner involvement in the learning task.	X	X	X			X	
III. Uses evaluation to improve instruction.							X
A. Uses information about learner performance to improve the instructional process.		X	X	X			
B. Reports learner status and progress to learners and their parents.	X	X	X	X		X	
IV. Manages classroom activities effectively.							X
A. Establishes and maintains appropriate learner behavior.	X	X		X		X	
B. Establishes and maintains a classroom climate conducive to learning.	X	X	X	X		X	
C. Makes effective use of classroom resources (e.g. personnel, time, materials, facilities)	X		X	X		X	
V. Establishes and maintains a professional leadership role.							
A. Improves professional skills and knowledge.		X	X	X			
B. Takes a leadership role in improving education.		X	X	X	X		
C. (Screening only) Performs professional responsibilities efficiently.			X	X	X		
VI. (Screening only) Communicates effectively.							X
A. Writes clearly and correctly.	X					X	
B. Communicates oral information effectively.	X		X	X		X	
C. Reads professionally relevant literature/materials with comprehension.		X	X	X			

BEST COPY AVAILABLE

HOW WE OBTAIN OUR INFORMATION

ABOUT

COMPETENCIES AND INDICATORS

GENERAL EDUCATION

	CLASSROOM OBSERVATION	DIALOGUE	PEER QUESTIONNAIRE	PRINCIPAL QUESTIONNAIRE	STUDENT QUESTIONNAIRE	PROFESSIONAL DEVELOPMENT AND LEADERSHIP SUBSKILLS	WRITTEN TEST
I. Prepares for instruction effectively.							X
A. Establishes appropriate instructional goals and related objectives consistent with the curriculum.		X		X			
B. Prepares instructional plans and materials incorporating principles of effective instruction.	X	X		X			
C. Creates, selects or modifies instructional plans and materials to accommodate learner instructional levels.	X	X		X			
II. Uses teaching strategies and procedures appropriate to the content, objectives, and learners.							X
A. Provides a clear description of the learning task and its content.	X	X		X	X		
B. Monitors learner understanding and reteaches as necessary.	X	X		X	X		
C. Provides learners appropriate practice and review.	X	X		X	X		
D. Establishes and maintains learner involvement in the learning task.	X	X		X	X		
III. Uses evaluation to improve instruction.							X
A. Uses information about learner performance to improve the instructional process.		X		X			
B. Reports learner status and progress to learners and their parents.	X	X		X	X		
C. Improves learner performance (Field Test Only)							
IV. Manages classroom activities effectively.							X
A. Establishes and maintains appropriate learner behavior.	X			X	X		
B. Establishes and maintains a classroom climate conducive to learning.	X	X		X	X		
C. Makes effective use of classroom resources (e.g. personnel, time, materials, facilities)	X			X	X		
V. Establishes and maintains a professional leadership role.							
A. Improves professional skills and knowledge.				X		X	
B. Takes a leadership role in improving education.			X	X		X	
C. (Screening only) Performs professional responsibilities efficiently.							
VI. (Screening only) Communicates effectively.							X
A. Writes clearly and correctly.							X
B. Reads professionally relevant literature/materials with comprehension.							X

BEST COPY AVAILABLE

Appendix B

List of Scores Used in Tennessee Career Ladder Evaluation for General Education Teachers

&

Analysis of Differentiation Among Career Ladder Levels in General Education 1985-86

&

Analysis of Effect Sizes for Differentiation of Levels in the Career Ladder Teacher System General Education 1985-86

BEST COPY AVAILABLE

List of Observation Scores

The following is a list of scores, with a verbal description of what they measure, and the items included in them. Pattern scores are identified whenever they occur. Also, some scores apply only to persons in a particular evaluation system. For instance, score T128 is used only in the evaluation of vocational education teachers. All "specialized" scores like T128 are identified in parentheses just after the score number.

T101. Fosters higher level thinking. Section 3, Planning/Strategies/Management, Teacher fosters higher level thinking skills.

T102. Teacher preparation. Section 3, Planning/Strategies/Management, Teacher preparation.

T103. Teaches task and content appropriately. Section 1, Teacher behavior, Task and content.

T104. Students ask procedural questions. Section 1, Procedural questions. NOTE: This is a pattern score.

T105. Students exhibit understanding. Section 2, question 4: Did students express/exhibit lack of understanding?

T106. Provides correct information. Section 2, question 6: Did the teacher present correct information?

T107. Communicates effectively. Section 3, Teacher communication, Provides an appropriate language model for learners.

T108. Monitors learner understanding. Section 1, Teacher behavior, Monitors.

T109. Section 2, question 5: Did the teacher respond to this by reteaching? This was a field-test score only. It was not used in your evaluation.

T110. Paces activities appropriately. Section 3, Planning/Strategies/Management, Teacher attention to needs in order of importance, and Teacher pacing of activities to accommodate learner understanding.

T111. Provides practice and review. Section 1, Teacher behavior, Practice and review.

T112. Students initiate/respond to lesson. Section 1, Student initiates/responds, Task, content, review, practice, monitors, feedback, and other - procedure, other - presentation.

BEST COPY AVAILABLE

T113. Students ask academic questions. Section 1, Academic questions. NOTE: This is a pattern score.

T114. Students are on task. Section 2, question 1: Students on task. This is calculated as the average difference between the number of students in class and the number of students off task, as coded in the grid at the top of Section 2.

T115. Teacher involves nonvolunteers. Section 1, Teacher involves students who do not volunteer.

T116. Student preparation. Section 3, Planning/Strategies/Management, Student preparation.

T117. Provides academic feedback. Section 1, Teacher behavior, Feedback.

T118. Provides reasonable sanctions. Section 2, question 8: Did the teacher provide reasonable sanctions for inappropriate behavior?

T119. Maintains appropriate student behavior. Section 1, Teacher control, Cause and result. NOTE: This is a pattern score.

T120. This score was collapsed into T119, as a part of the pattern score.

T121. Positive classroom climate. Section 1, Positive affect (Task, content, review, practice, monitors, feedback, other) and Teacher behavior - other, greeting/socialization.

T122. Students treated without favoritism. Section 3, Favoritism.

T123. Maximizes available learning time. Section 1, Time elapsed between when the interruption ended and the class was back on task.

T124. Uses classroom resources effectively. Section 3, Use of resources.

T125. Uses classroom facilities effectively. Section 3, Facilities.

T126. Engages in classroom procedures. Section 1, Teacher behavior, Other - classroom procedure.

T127. Teacher behavior is on task. Section 1, Teacher behavior, Other - off-task.

BEST COPY AVAILABLE

DIALOGUE

After meeting with the teacher and having a focused discussion about planning, teaching strategies, or evaluation, the evaluator assigned ratings (on a 1 - 5 scale) to a series of measurement statements keyed to particular competencies. The Dialogue instrument is shown on pp. 84-93 of your Orientation Manual. Your scores are the average ratings given on the specified items. Check Table 5 of Appendix A or your score report to determine which of these scores was used in your evaluation.

T201. Goals/objectives address the core curriculum. Planning focus, Items 1 through 4.

T202. Uses principles of effective instruction. Planning focus, Items 5 through 8.

T203. Plans accommodate student differences. Planning focus, Items 9 through 11.

T204. Describes learning task and content clearly. Teaching strategies focus, Items 1 through 5.

T205. Attends to learner needs and differences. Teaching strategies focus, Items 6 through 9.

T206. Provides for appropriate review and practice. Teaching strategies focus, Items 10 through 13.

T207. Creates and maintains learner involvement. Teaching strategies focus, Items 14 through 17.

T208. Uses assessment to improve teaching. Evaluation focus, Items 1 through 4, and 6.

T209. Monitors and reports student progress. Evaluation focus, Items 5, 7 and 8.

T210. Evaluation focus, Items 11 and 12. This was a field-test score which was not used in your evaluation.

T211. Supports learner achievement. Evaluation focus, Items 9 and 10.

BEST COPY AVAILABLE

PEER QUESTIONNAIRE

Three of the six colleagues whose names you submitted responded independently to this instrument. It consists of 16 items (18 for Special Populations) which are answered in terms of frequency, from "Rarely" to "Almost always." See pages 73-74 of the Orientation Manual for a copy of this instrument.

T301. Maintains a leadership role with others. Items 1 through 16.

PRINCIPAL QUESTIONNAIRE

The scores from this instrument are listed last because they are not added in to all the other scores in the data sources component. This cannot be done because the scores on the Principal Questionnaire are not expressed as standard scores. Instead, your principal responded to the items on this instrument directly on the 200-800 scale (unlike the description given in the Orientation Manual, on pp. 82-83). A handout explaining the changes in the response scale was sent to your principal along with the instrument (see Appendix D for a copy). The weights assigned to the principal questionnaire have been "pulled out" of the data source weights reported in your Teacher Orientation Manual, and have been added back in as a separate component. This does not change the weight given to the principal questionnaire.

On the principal questionnaire three scores were used to judge indicators that were considered minimum competency. A passing value for each of these scores was a 450 or above. As with other screening sources (reading and writing), a pass was required to be awarded certification for the upper levels of Career Ladder.

T401. Goals/objectives tied to curriculum. Item 1.

T402. Plans for effective instruction. Item 2.

T403. Plans accommodate learner needs. Item 3.

T404. Communicates task and content effectively. Items 4 and 17.

T405. Adapts lesson for learner understanding. Item 5.

T406. Provides review and practice as needed. Item 6.

T407. Creates and maintains academic involvement. Item 7.

T408. Uses assessment to enhance teaching. Item 8.

BEST COPY AVAILABLE

T409. Reports progress to learners and parents. Item 9.

T410. Items 19 and 20. This was only a field-test score. It was not used in your evaluation.

T411. Maintains appropriate learner behavior. Item 10.

T412. Maintains climate conducive to learning. Item 11.

T413. Classroom resources used effectively. Item 12.

T414. Improves professional skills and knowledge. Item 13.

T415. Improves education via leadership role. Item 14.

STUDENT QUESTIONNAIRES

Three levels of student questionnaires were available: primary, elementary, and secondary. Only one level of student questionnaire was used in your evaluation--the elementary version, the secondary version, or the primary version, depending upon the age and/or comprehension level of the students you teach. The individual levels are described below.

Student Questionnaire - Elementary

The items on this instrument are shown on pp. 76-78 of your Orientation Manual. Students were asked to respond to each item on a 5-point scale, where "Almost never" = 1, "Sometimes" = 2, "About half the time" = 3, "Usually" = 4, and "Almost always" = 5. Nine scores are derived from the average responses of your students to the 34 items. Score order is reversed (i.e., "Almost never" = 5; "Almost always" = 1) for some items, as noted. Also, the Special Populations version of this questionnaire used a 3-point response scale ("Never" = 1, "Some of the time" = 2, "Almost always" = 3).

T501. Explains academic task to learner. Item 1.

T502. Paces instruction appropriately. Items 2, 3, 24, and 25 (items 24 and 25 scored in reverse order).

T503. Provides opportunity to review/practice. Items 6, 8, 9, 10, 11, 13, 14, and 22.

T504. Reinforces importance of schoolwork. Items 5 and 7.

BEST COPY AVAILABLE

T505. Assists learners through feedback. Items 12, 15, 16, and 17.

T506. Items 33 and 34. This was a field-test score, so it was not used in your evaluation.

T507. Manages learner behavior effectively. Items 18 through 21.

T508. Encourages achievement by all learners. Items 27 through 32 (item 32 scored in reverse order).

T509. Utilizes resources effectively. Items 4, 23, and 26.

Student Questionnaire - Secondary

Your students responded to 39 items on a 5-point scale, from "Rarely" = 1, "Sometimes" = 2, "About half the time" = 3, "Usually" = 4, to "Almost always" = 5. The items are shown on pp. 79-81 of the Teacher Orientation Manual. The ten scores derived from this instrument equal the average ratings your students marked on pertinent items. Again, the form used with Special Populations had a 3-point scale, from "Never" = 1, "Some of the time" = 2, to "Almost always" = 3.

T601. Explains academic task and content. Items 1 and 2.

T602. Paces instruction appropriately. Items 3, 30 and 31 (items 30 and 31 scored in reverse order).

T603. Encourages appropriate review/practice. Items 4, 9, 10, 12 through 15, and 28.

T604. Reinforces importance of schoolwork. Items 7 and 8.

T605. Assists learners through feedback. Items 11, 16 through 19, 22 and 23.

T606. Items 38 and 39. This was a field-test score that was not used in your evaluation.

T607. Manages learner behavior effectively. Items 21, 25 through 27.

T608. Encourages achievement by all learners. Items 20, 33 through 37 (item 37 scored in reverse order).

T609. Utilizes resources effectively. Items 5, 6, 24, 29, and 32 (item 24 scored in reverse order).

BEST COPY AVAILABLE

Student Questionnaire - Primary

Your students responded to the nine items on this questionnaire by using a "sign" system established by the evaluator, for example, thumbs up to signify "agree," thumbs down to signify "disagree," and arms folded to indicate "can't say" or "don't know." The evaluator recorded the responses of each student. "Yes" was coded as "2" and "No" as "1." The yes/no responses were averaged for each of the three scores from this instrument (see page 75 of your Orientation Manual for a copy of the statements).

T701. Uses appropriate instructional strategies. Items 1, 2, and 5.

T702. Reports and monitors student progress. Items 3, 8, and 9.

T703. Effectively manages activities. Items 4, 6, and 7.

PROFESSIONAL SKILLS TEST

This part of the Tennessee Career Ladder Test consists of 100 multiple-choice items in four sections which correspond to the domains of Planning, Teaching Strategies, Evaluation, and Classroom Management. Each score equals the number of correct responses made out of 25 possible for the particular section. This instrument, unlike all the others, is not printed in your Orientation Manual. It is a secure test, by order of the Certification Commission. You may wish to examine your Study Guide to remind you of the types of items included on each portion of the test.

T801. Knowledge of planning. Items 1 through 25.

T802. Knowledge of teaching strategies. Items 26 through 50.

T803. Knowledge of evaluation. Items 51 through 75.

T804. Knowledge of classroom management. Items 76 through 100.

BEST COPY AVAILABLE

PROFESSIONAL DEVELOPMENT AND LEADERSHIP SUMMARY

Five categories of development activities and five categories of leadership activities were rated independently by two evaluators on a 5-point scale, from "unsatisfactory" (1), "below average" (2), "average" (3), "outstanding" (4), to "distinguished" (5). The two evaluators came to a consensus on each of the ten items. If a consensus could not be reached, a third evaluator was consulted. The ten items were broken down into two sets of five ratings to form two scores. As with the other instruments, scores which are specialized to particular types of teachers are identified as such in parentheses just after the score number.

T901. Enhances instruction with new techniques. Categories 1 through 5.

T902. Exhibits leadership to improve schooling. Categories 6 through 10.

CONSENSUS RATING

At the end of the evaluation process the A, B, and C evaluators for each candidate met to discuss their professional judgment of each candidate. This judgment was made for the domains of planning, teaching strategies, evaluation, and classroom management. A consensus judgment was not made in leadership. The rating was made on the same direct judgment scale used by the principals. See page 80 of Appendix D for this scale. The consensus judgment was given a weight of 25% for each of the four corresponding domains in which a judgment was made.

BEST COPY AVAILABLE

Table 1:

ANALYSIS OF DIFFERENTIATION AMONG CAREER LADDER LEVELS IN GENERAL EDUCATION 1985-86 DATA

Score	Average Score for Level I	II	III	Standard Deviation	Differentiates Between Levels I&II	I&III	II&III	Total N
N-about	132	307	629					
T101	50.06	60.75	73.52	14.35	*	*	*	1068
T102	62.18	76.91	89.61	13.08	*	*	*	1068
T103	10.39	11.95	13.72	4.38	*	*	*	1068
T104	72.72	78.68	82.94	17.49	*	*	*	1068
T105	87.97	88.39	90.20	7.18	-	*	*	1068
T106	99.71	99.80	99.92	.85	-	*	-	1068
T107	88.24	93.09	97.03	7.49	*	*	*	1068
T108	14.62	18.01	22.38	6.46	*	*	*	1068
T110	55.94	68.94	82.70	13.05	*	*	*	1068
T111	7.82	8.51	10.03	3.03	-	*	*	1068
T112	7.96	8.87	10.11	1.94	*	*	*	1068
T113	61.63	66.27	69.08	18.57	*	*	-	1068
T114	99.93	99.98	99.99	.04	*	*	*	1068
T115	49.00	56.48	64.14	17.75	*	*	*	1068
T116	56.09	67.45	79.23	13.15	*	*	*	1068
T117	16.26	19.52	22.49	5.79	*	*	*	1068
T118	67.02	86.72	94.11	23.04	*	*	*	876
T119	98.82	99.58	99.83	.65	*	*	*	1068
T121	1.03	1.32	2.08	.99	*	*	*	1068
T122	100.00	99.95	100.00	.51	-	-	-	1068
T123	99.84	99.84	99.93	.16	-	-	-	77
T124	55.92	67.41	79.91	12.18	*	*	*	1068
T125	59.46	68.22	79.16	13.52	*	*	*	1068
T126	85.72	87.02	88.14	10.33	-	*	-	1068
T127	98.18	99.41	99.84	1.62	*	*	*	1068
T201	3.11	3.58	4.22	.72	*	*	*	1068
T202	3.07	3.57	4.21	.74	*	*	*	1068
T203	2.80	3.24	3.90	.81	*	*	*	1068
T204	3.24	3.75	4.22	.60	*	*	*	1068
T205	2.81	3.43	3.99	.70	*	*	*	1068
T206	2.93	3.52	4.10	.70	*	*	*	1068
T207	2.96	3.54	4.15	.70	*	*	*	1059
T208	2.80	3.36	4.00	.79	*	*	*	1067
T209	2.92	3.38	4.06	.80	*	*	*	1067
T210	2.41	2.88	3.41	.93	*	*	*	1066
T211	2.62	3.22	3.87	.85	*	*	*	1067
T301	4.59	4.76	4.86	.25	*	*	*	1068
T501	4.46	4.58	4.73	.30	-	*	*	256
T502	4.03	4.29	4.45	.32	*	*	*	256
T503	3.43	3.74	4.02	.46	*	*	*	256

Table 1 cont'd

T504	4.03	4.21	4.37	.36	-	*	*	256
T505	3.52	3.74	4.07	.47	-	*	*	256
T506	4.29	4.50	4.67	.34	*	*	*	256
T507	4.10	4.29	4.50	.34	*	*	*	256
T508	4.15	4.34	4.54	.31	*	*	*	256
T509	4.03	4.32	4.54	.36	*	*	*	256
T601	4.28	4.54	4.69	.33	*	*	*	619
T602	4.28	4.46	4.57	.29	*	*	*	619
T603	3.47	3.75	3.97	.44	*	*	*	619
T604	3.78	4.01	4.21	.40	*	*	*	619
T605	3.37	3.66	3.94	.46	*	*	*	619
T606	4.01	4.31	4.51	.42	*	*	*	619
T607	3.74	3.90	4.03	.36	*	*	*	619
T608	4.11	4.32	4.54	.35	*	*	*	619
T609	4.36	4.54	4.68	.27	*	*	*	619
T701	1.96	1.95	1.98	.04	-	-	*	193
T702	1.89	1.94	1.96	.06	*	*	-	193
T703	1.96	1.99	1.99	.03	*	*	-	193
T901	2.62	2.98	3.29	.68	*	*	*	1068
T902	2.37	2.75	3.09	.68	*	*	*	1067
T801	19.26	20.25	21.07	2.72	*	*	*	1068
T802	18.44	19.50	20.07	2.41	*	*	*	1068
T803	15.74	16.89	17.39	3.24	*	*	-	1068
T804	19.72	20.93	21.47	2.67	*	*	*	1068
Princpl 1	739.91	770.24	785.11	41.53	*	*	*	1068
Princpl 2	733.61	771.31	785.10	43.40	*	*	*	1068
Princpl 3	734.14	763.87	779.09	44.72	*	*	*	1068
Princpl 4	727.25	768.63	784.33	51.14	*	*	*	1068
Princpl 5	736.53	762.04	780.03	47.63	*	*	*	1068
Consens 1	494.98	641.60	734.91	104.38	*	*	*	1068
Consens 2	459.66	626.93	726.33	113.02	*	*	*	1068
Consens 3	461.59	619.45	719.33	112.19	*	*	*	1068
Consens 4	453.01	621.51	729.39	125.68	*	*	*	1068
Datadom 1	459.54	584.14	719.73	125.10	*	*	*	1068
Datadom 2	495.34	661.58	723.39	88.31	*	*	*	1068
Datadom 3	555.67	668.12	726.97	85.22	*	*	*	1068
Datadom 4	543.04	682.39	735.24	85.12	*	*	*	1068
Datadom 5	648.33	714.12	738.94	63.17	*	*	*	1068
Scoredm 1	496.44	617.11	730.06	101.97	*	*	*	1068
Scoredm 2	495.95	657.31	726.59	85.34	*	*	*	1068
Scoredm 3	550.00	665.53	730.27	77.86	*	*	*	1068
Scoredm 4	542.63	677.52	739.66	78.99	*	*	*	1068
Scoredm 5	665.97	723.71	747.16	54.31	*	*	*	1068

ANALYSIS OF EFFECT SIZES FOR DIFFERENTIATION OF LEVELS IN THE CAREER LADDER TEACHER SYSTEM GENERAL EDUCATION -- 1985-6 DATA

Score	Weight	Difference Between Levels In Standard Deviation Units Level I & II	Level II & III
(Observation)			
T101	.0195	.74	.89
T102	.0105	1.13	.97
T103	.0128	.36	.40
T104	.0052	.34	.24
T105	.0030	.06	.25
T106	.0030	.14	1.00
T107	.0015	.65	.53
T108	.0128	.52	.68
T110	.0064	1.00	1.05
T111	.0240	.23	.50
T112	.0098	.47	.64
T113	.0098	.25	.15
T114	.0052	1.25	.25
T115	.0075	.42	.43
T116	.0052	.86	.90
T117	.0300	.56	.51
T118	.0090	.86	.32
T119	.0270	1.17	.39
T121	.0240	.29	.77
T122	.0120	-.10	.10
T123	.0038	0.00	.56
T124	.0075	.94	1.03
T125	.0030	.65	.81
T126	.0019	.13	.11
T127	.0019	.59	.26
(Dialogue)			
T201	.0240	.65	.90
T202	.0180	.68	.86
T203	.0135	.54	.82
T204	.0210	.85	.78
T205	.0210	.89	.80
T206	.0210	.84	.83
T207	.0120	.83	.87
T208	.0240	.71	.81
T209	.0165	.57	.85
T211	.0075	.71	.76

Score	Weight	Difference Between Levels In Standard Deviation Units Level I & II	Level II & III
(Peer Questionnaire)			
T301	.0300	.68	.40
(Elementary/Secondary Questionnaire)			
T501 / T601	.0075	.40 / .79	.50 / .45
T502 / T602	.0075	.81 / .62	.50 / .38
T503 / T603	.0090	.67 / .64	.61 / .50
T504 / T604	.0090	.50 / .58	.44 / .50
T505 / T605	.0150	.47 / .63	.70 / .61
T507 / T607	.0135	.56 / .44	.62 / .36
T508 / T608	.0165	.61 / .60	.64 / .63
T509 / T609	.0105	.80 / .67	.61 / .52
(Primary Questionnaire)			
T701	.0330	-.25	.75
T702	.0150	.83	.33
T703	.0405	1.00	0.00
(Professional Skills Test)			
T801	.0120	.36	.30
T802	.0270	.44	.24
T803	.0120	.35	.15
T804	.0195	.45	.20
(Professional Development & Leadership Summary)			
T901	.0240	.53	.46
T902	.0260	.56	.50
(Principal Questionnaire)			
Principal 1	.0150	.73	.36
Principal 2	.0150	.87	.32
Principal 3	.0150	.66	.34
Principal 4	.0300	.81	.31
Principal 5	.0200	.54	.38
(Evaluator Consensus Judgment)			
Consensus 1	.0375	1.40	.89
Consensus 2	.0875	1.48	.88
Consensus 3	.0375	1.41	.89
Consensus 4	.0625	1.34	.86